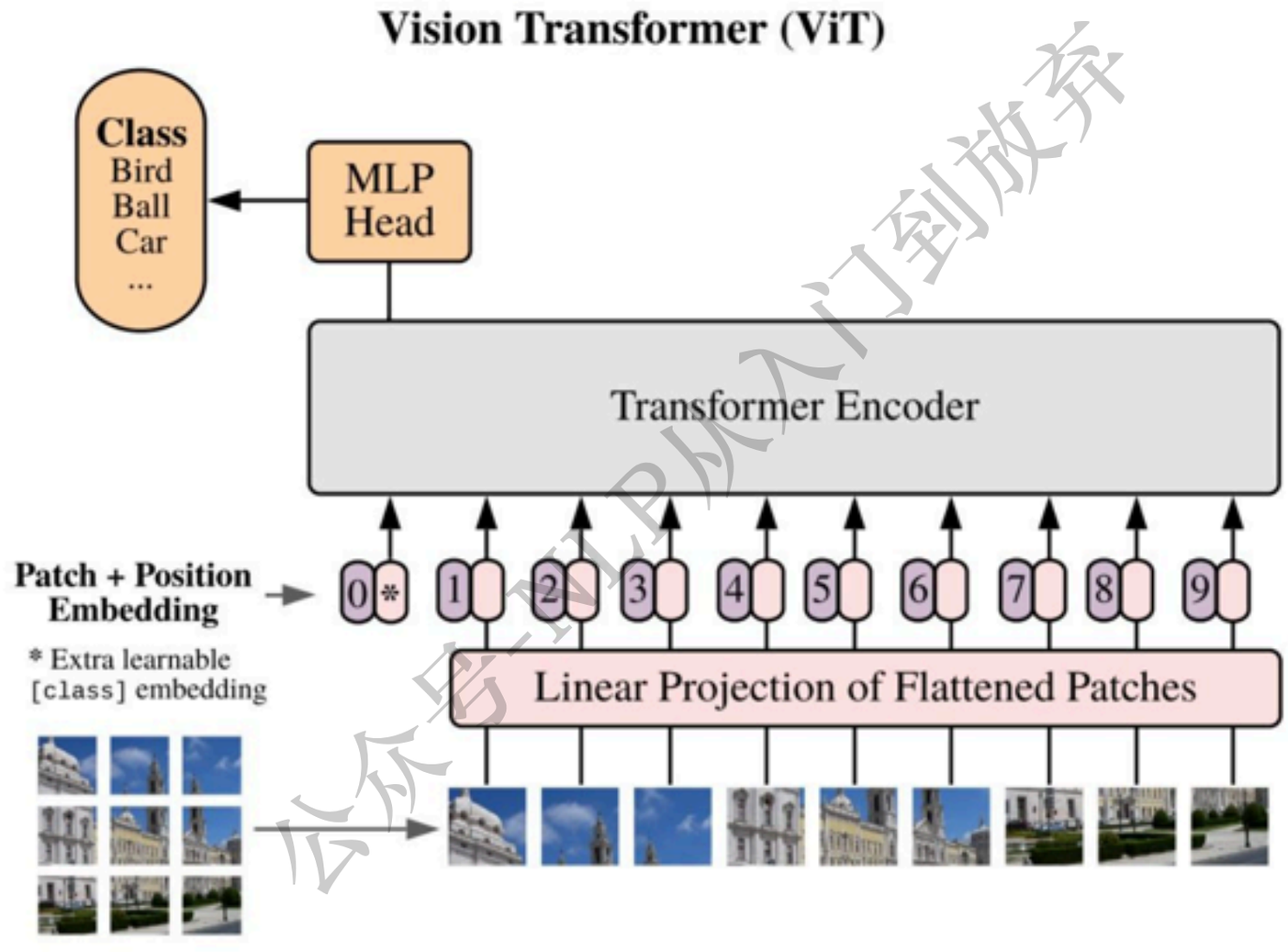
Vision Transformer 从零解读



后台回复【VIT】获取对应的代码和PPT

扫码关注微信公众号

文章周更

知识分享

一起进步

求关注，求点赞，求一切！！

TRM模型架构图

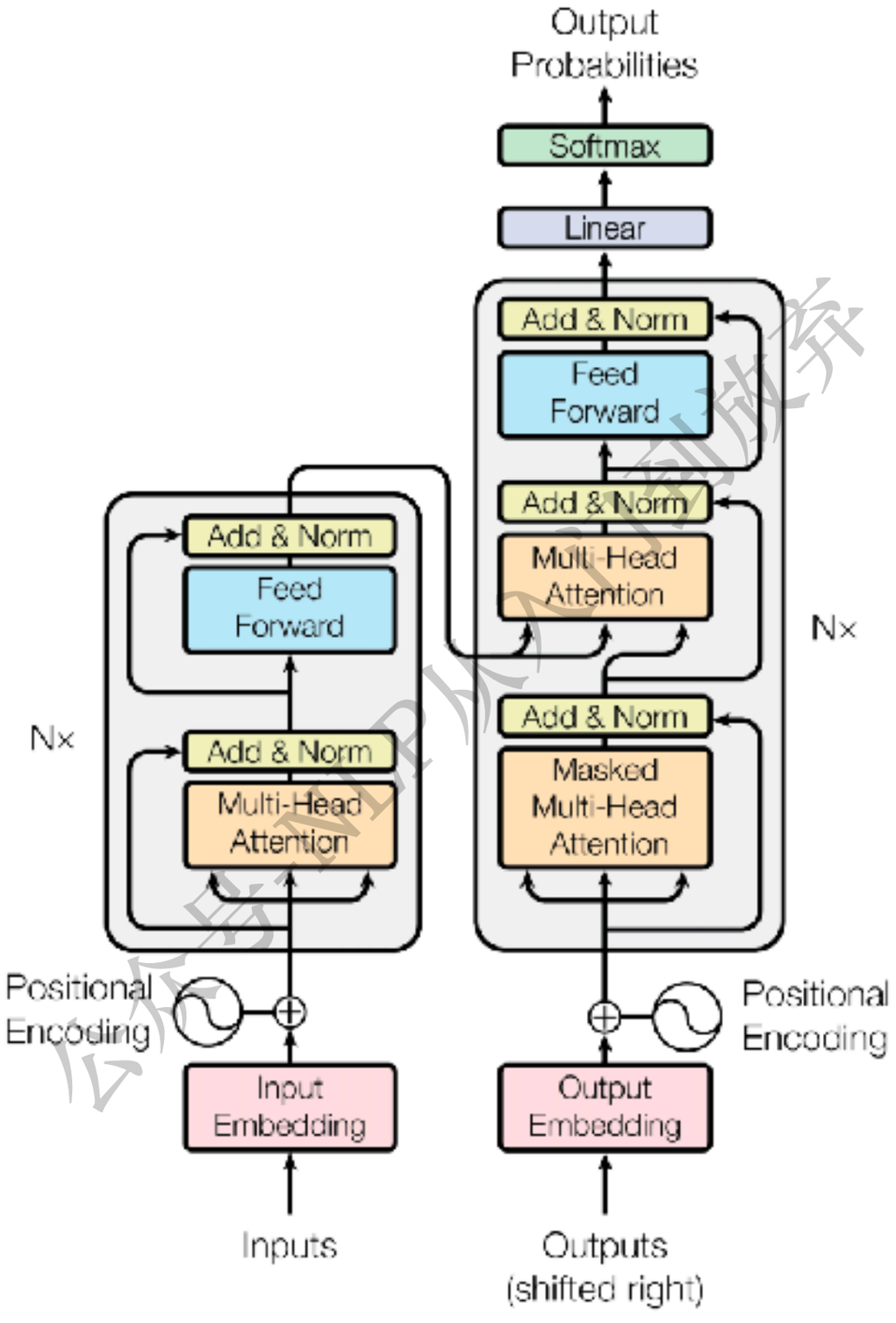
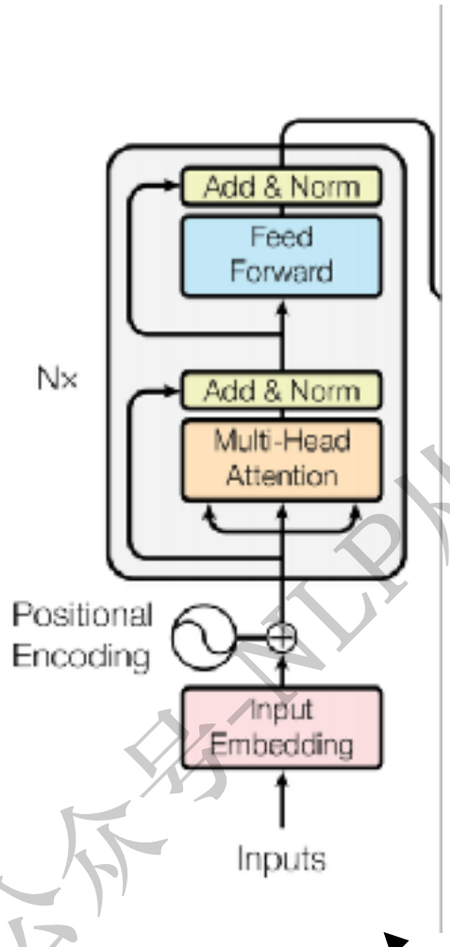


Figure 1: The Transformer - model architecture.



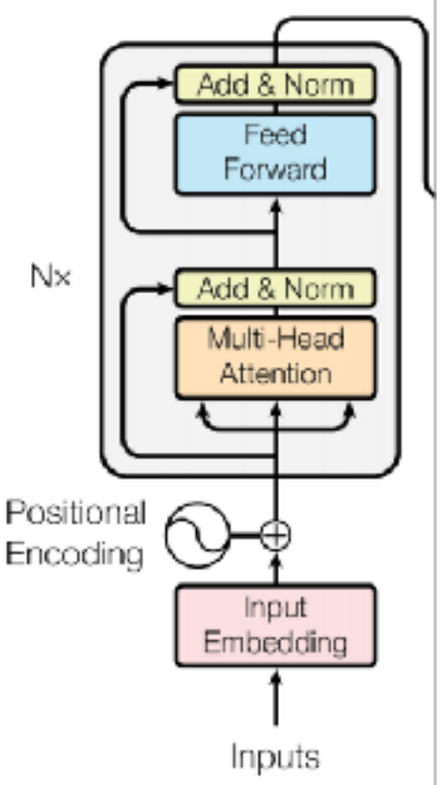
字符转化为数字

原始字符

3 5 7

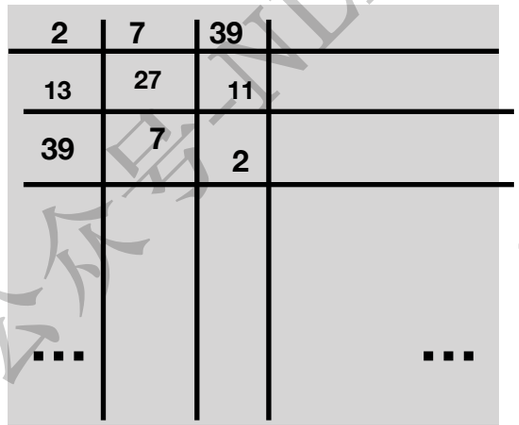
我 爱 你

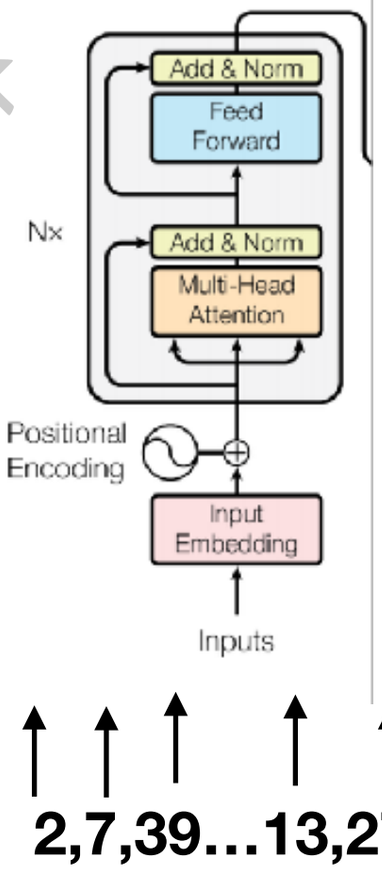
如何把图片融入到TRM中去



大部分人的思路

224





2,7,39...13,27,11...

复杂度的问题

224*224*1

224

2	7	39	
13	27	11	
39	7	2	
...			...

224

→ 序列长度=224*224=50176

BERT的最大长度是512，相当于100倍

如何处理复杂度的问题？：本质上是去解决随着像素增加，复杂度平方级增长的问题；

1.局部注意力机制

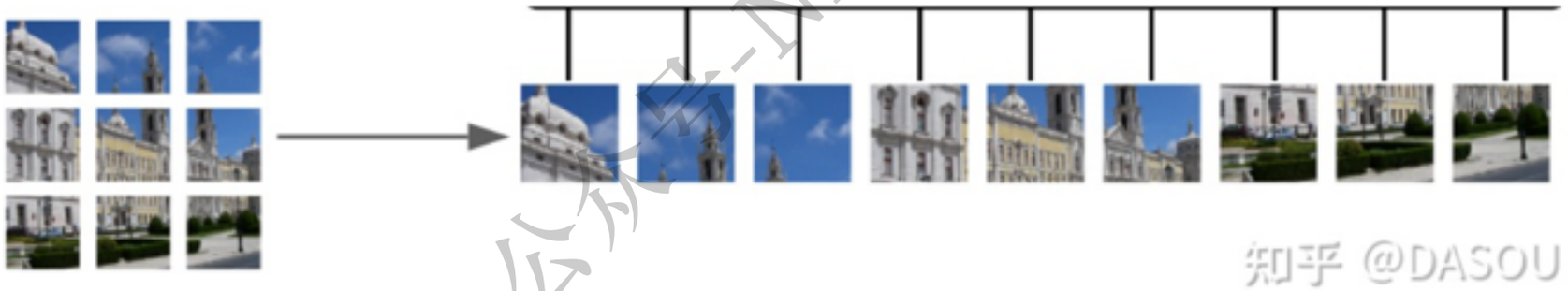
有很多中方法：

2. 改进attention公式

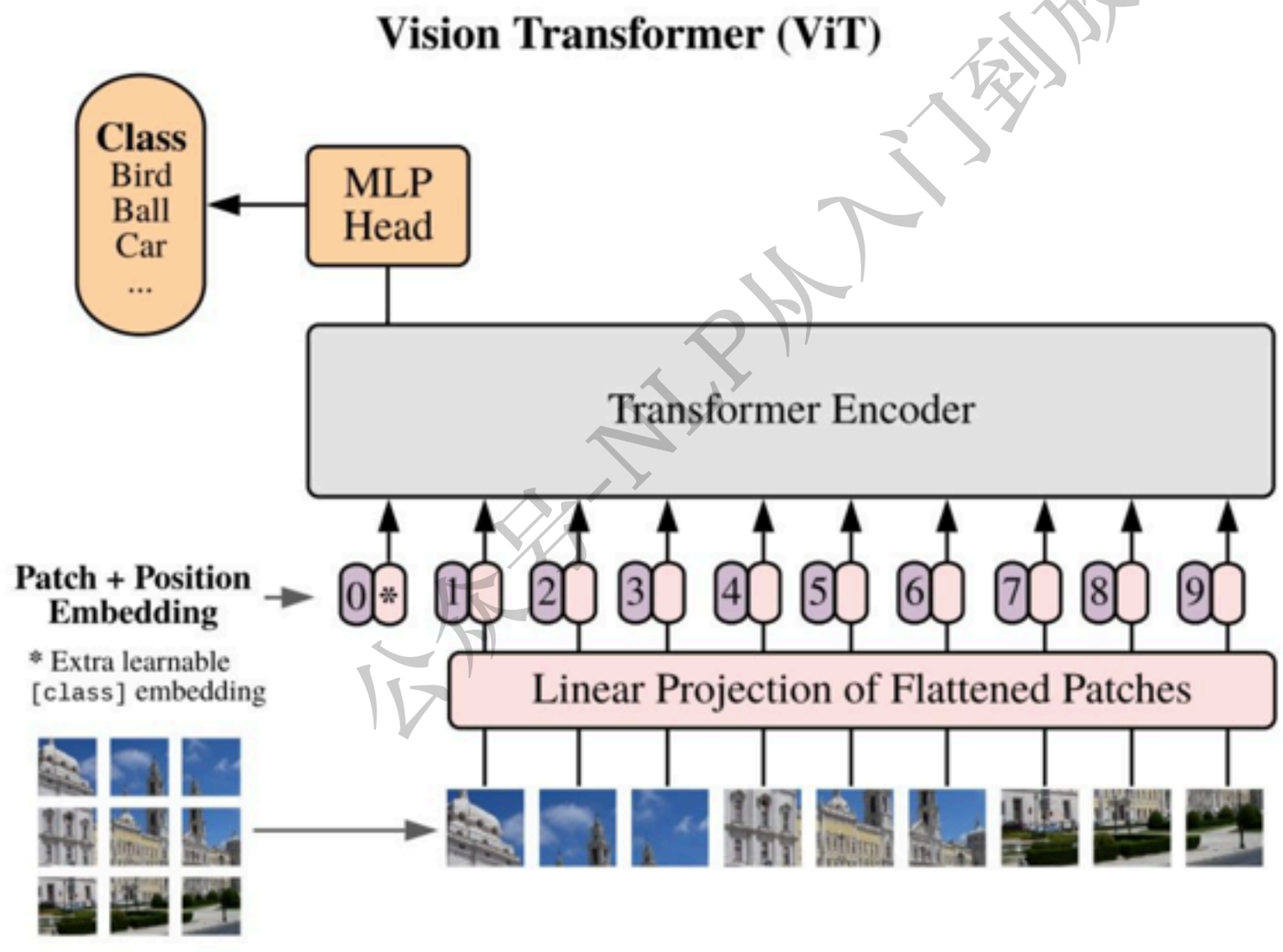
3……..

一个简单的改进方式：图像化整为零，切分patch

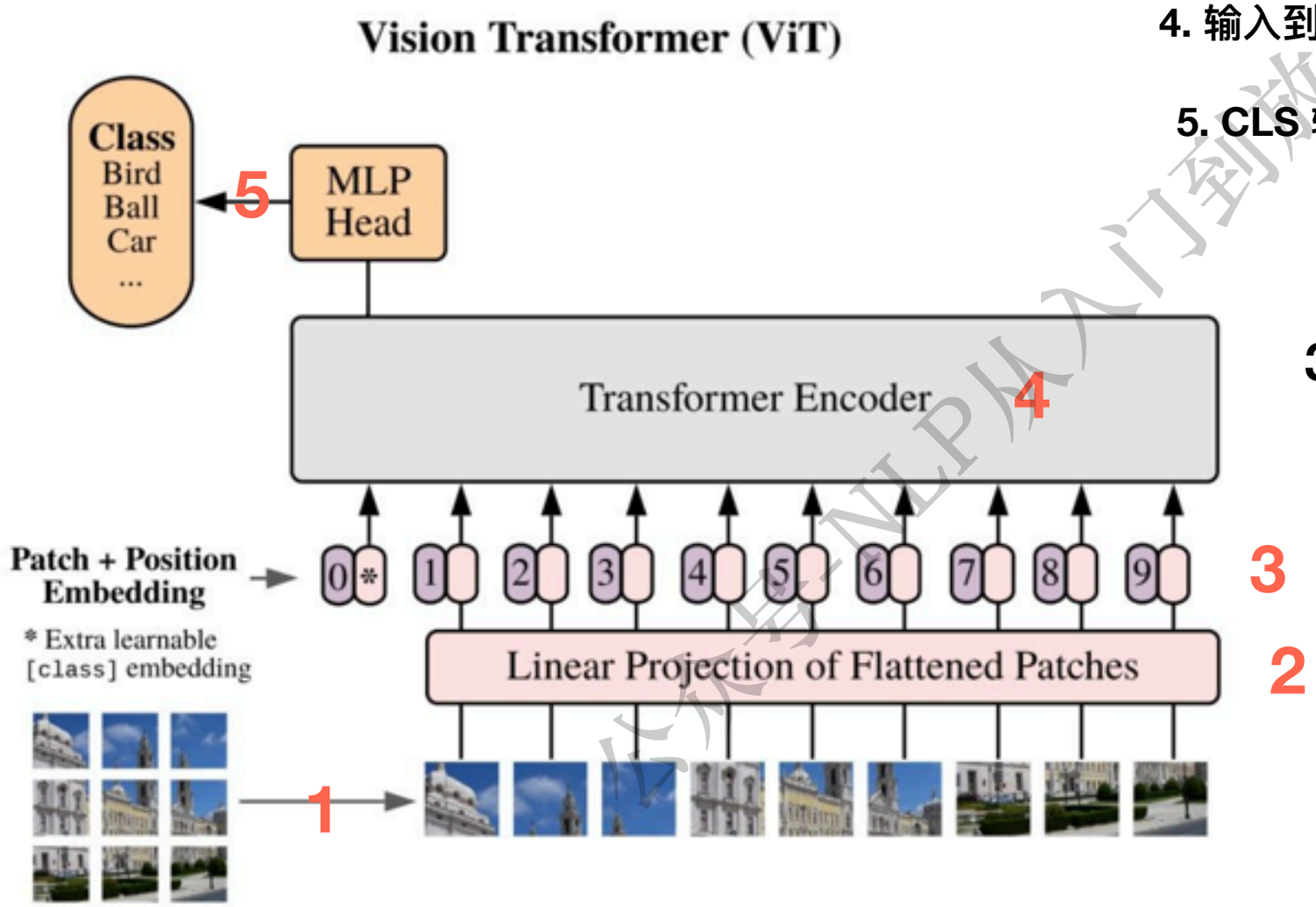
也就说原来是一个像素点代表一个token，
现在是一大块的token一个patch作为一个token



VIT模型架构图：



VIT模型架构图：



1. 图片切分为patch

2. patch转化为embedding

3.位置embedding和tokensembedding相加

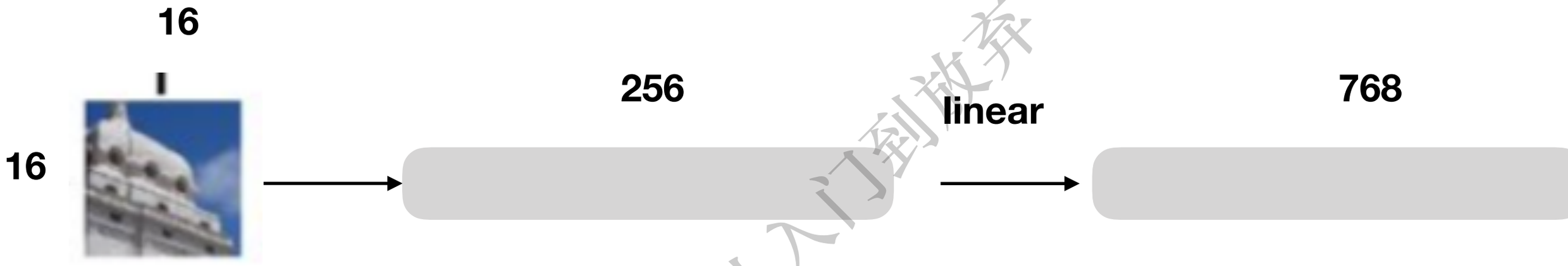
4. 输入到TRM模型

5. CLS 输出做多分类任务

3.1 生成CLS符号的TokenEMB

3.2 生成所有序列的位置编码

3.3 tonken+位置编码



为什么加入一个CLS符号

原论文中是这样说的：

In order to stay as close as possible to the original Transformer model, we made use of an additional [class] token, which is taken as image representation.

在整合最后输出信息的时候，有多种方式

两种方式，一种是使用【CLS】token，另一种就是对所有tokens的输出做一个平均

BERT有两个预训练任务

1.NSP任务：预测下一句

2. MLM：预测当前单词

BERT为什么采用一个CLS符号呢?

我自己的猜测是：如果采用一个平均，会涉及到所有tonkens的输出；
而MLM任务又会涉及到其中的部分mask的tokens的输出；

CSL符号一定程度在让两个任务保持一种相对的独立；

但是VIT不涉及到MLM这种形式的任务，只会有一个多分类任务，所以CLS符号不是必须的

在整合图片信息的时候，两种方式，一种是使用【CLS】token，另一种就是对所有tokens的输出做一个平均，简称GAP；实验结果证明，两者可以达到的同样的效果，只不过要控制好学习率；

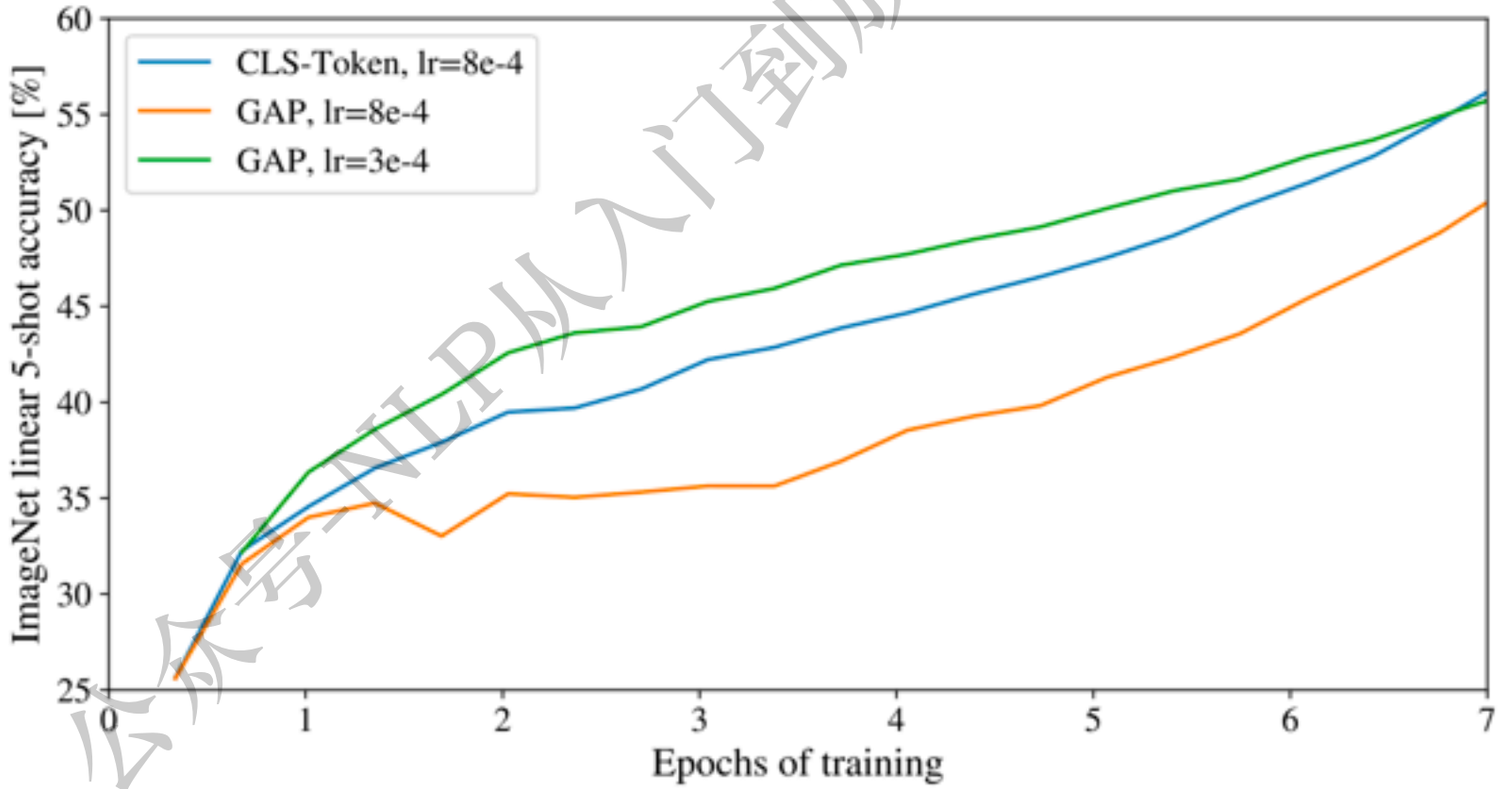


Figure 9: Comparison of class-token and global average pooling classifiers. Both work similarly well, but require different learning-rates.

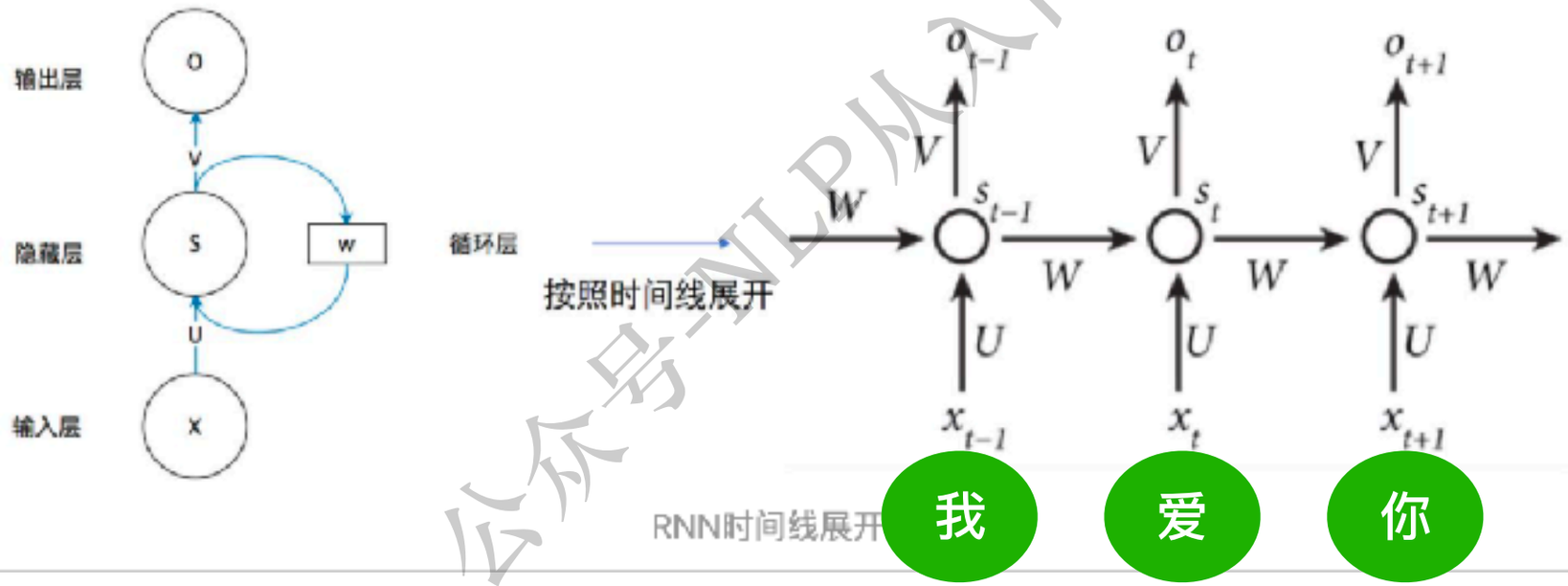
位置编码

1. 为什么需要位置编码

2. 为什么位置编码可以和patch embeding相加

很多CV出身的朋友不了解位置编码的重要性

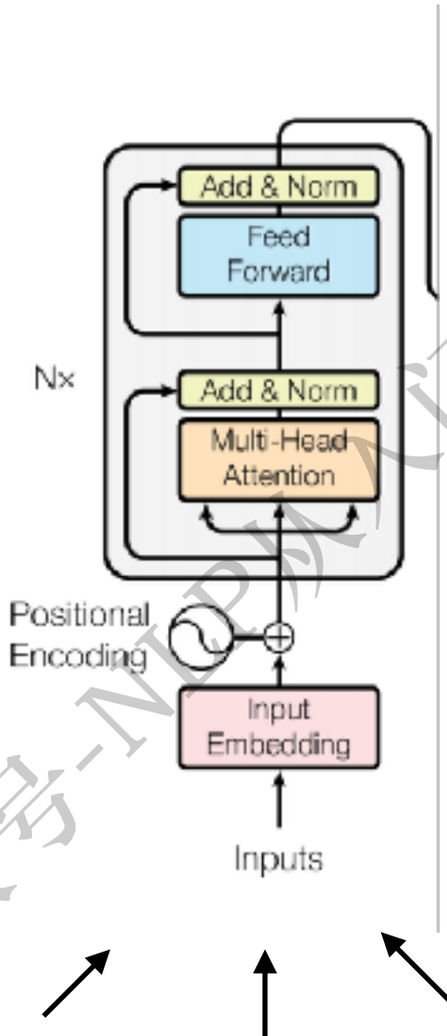
在我之前的TRM讲解视频，对于位置编码视频重要性的讲解是从RNN引出来的



接受上一个时刻的隐层信息，一个个的运算，天然的时序关系

TRM不是这样的：编码器天然并行，所有词汇一起输入；

不存在等待之前单词输出信息的情况



所以需要一个位置编码

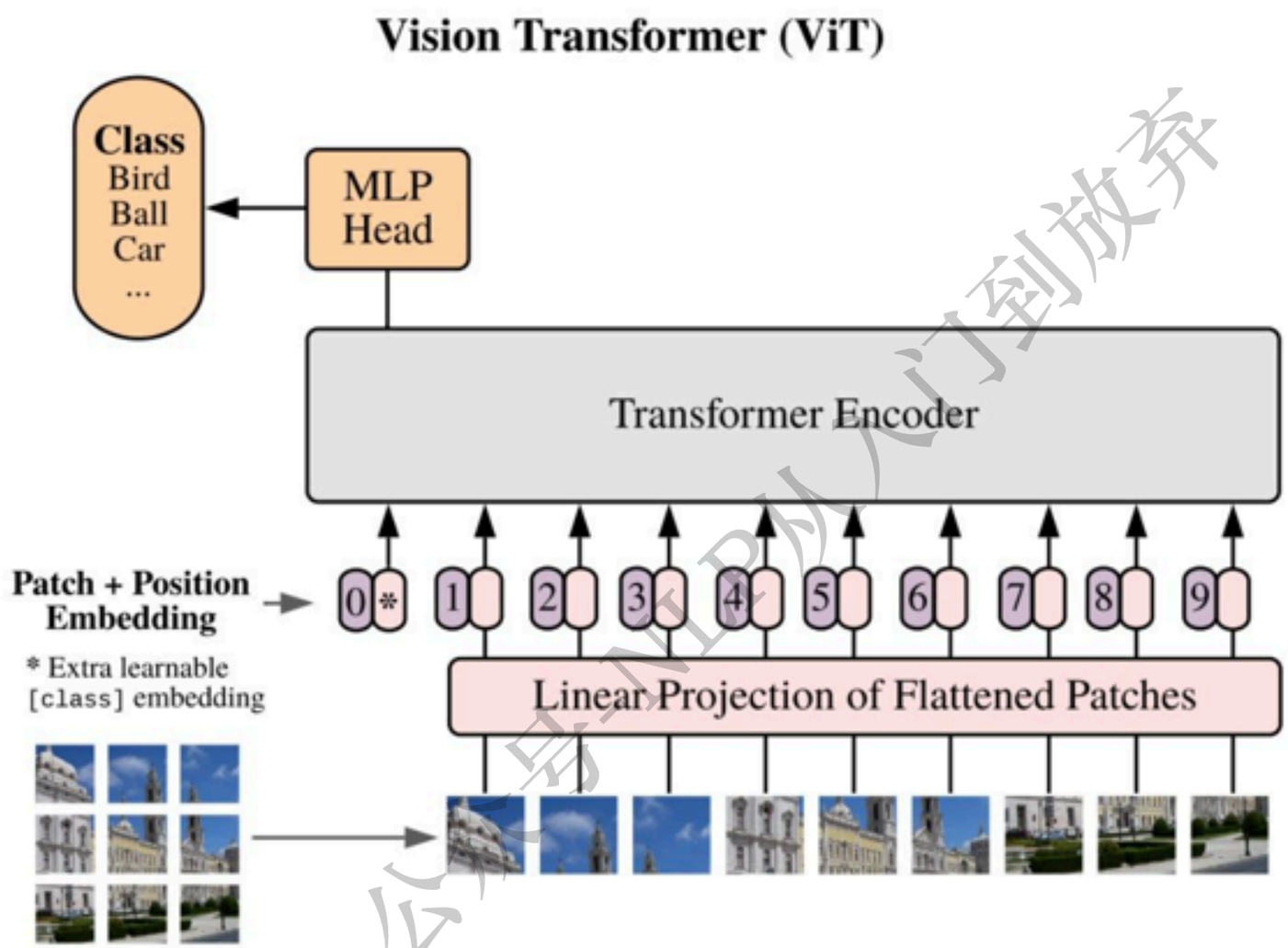
字符转化为数字

3 5 7

原始字符

我 爱 你

VIT中的位置编码：可学习的参数

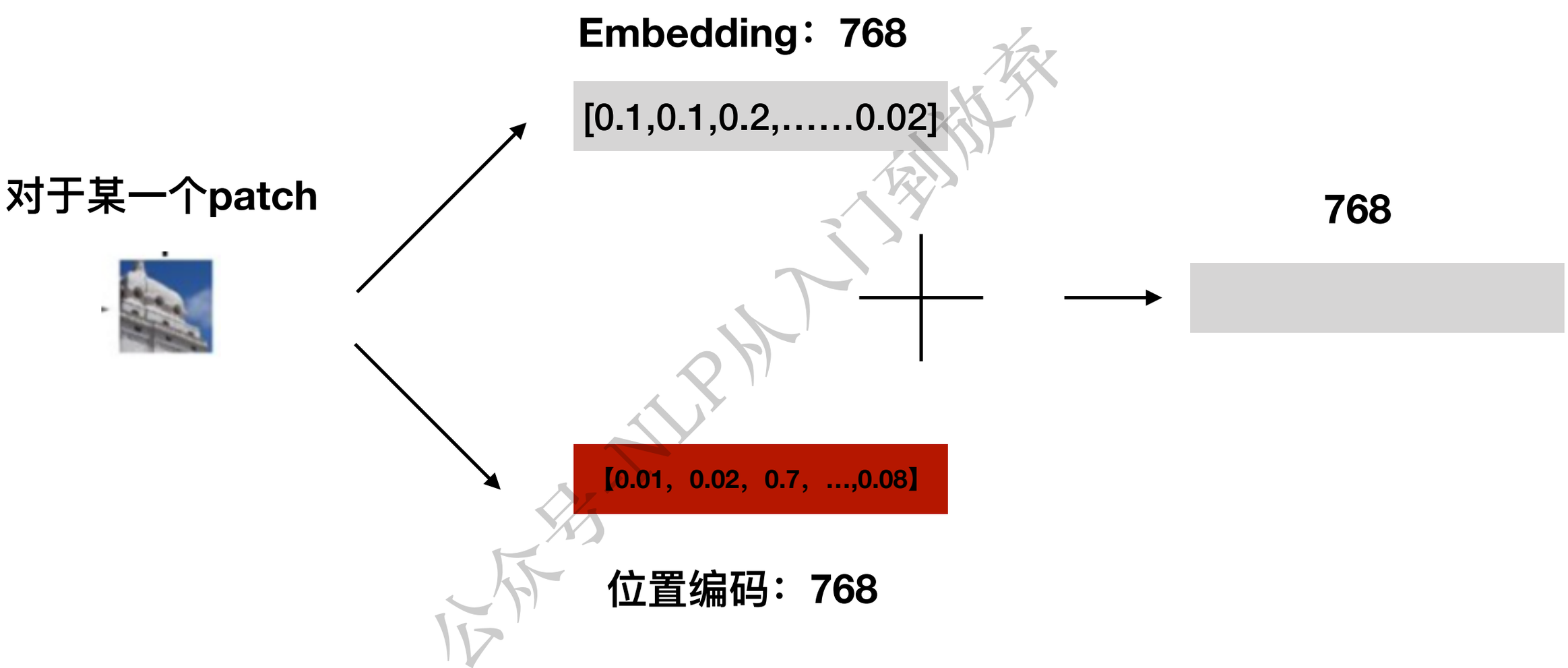


1 2 3 4 5 6 7 8 9 一维

【1，1】【1，2】 ..... 二维

相对位置信息

位置编码-CV



为什么Patch embedding和位置编码可以相加?

我看过很多解释，大家都在以果推因；
就是看到这个模型是这样做的，然后去推断这么做的原因；
记住就好；

最开始

每一层都加入
而且独立训练

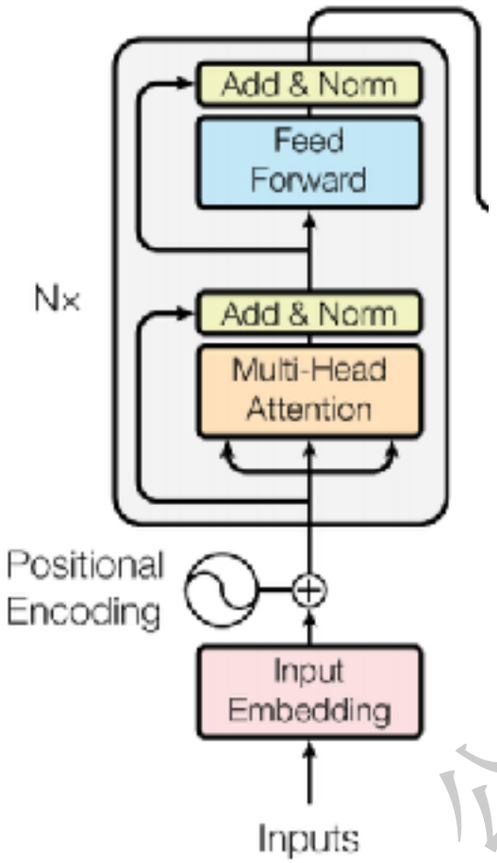
每一层都加入
但是参数共享

Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
没有位置编码 No Pos. Emb.	0.61382	N/A	N/A
一维位置编码 1-D Pos. Emb.	0.64206	0.63964	0.64292
二维位置编码 2-D Pos. Emb.	0.64001	0.64046	0.64022
相对位置编码 Rel. Pos. Emb.	0.64032	N/A	N/A

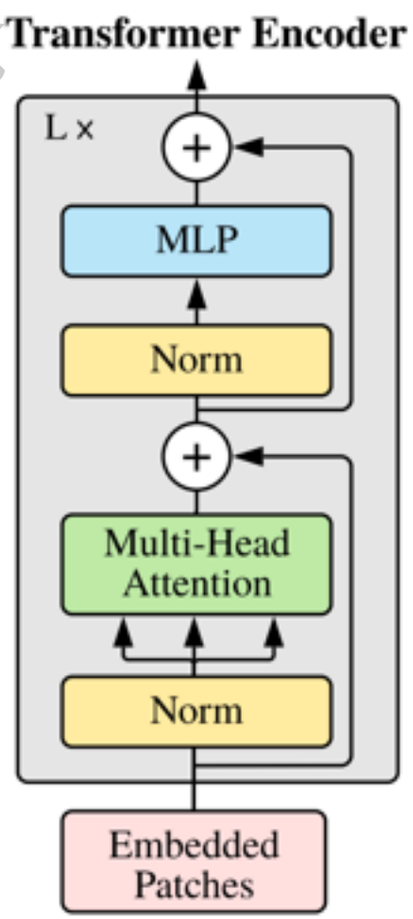
Table 8: Results of the ablation study on positional embeddings with ViT-B/16 model evaluated on ImageNet 5-shot linear.

TRM编码部分：

原始：

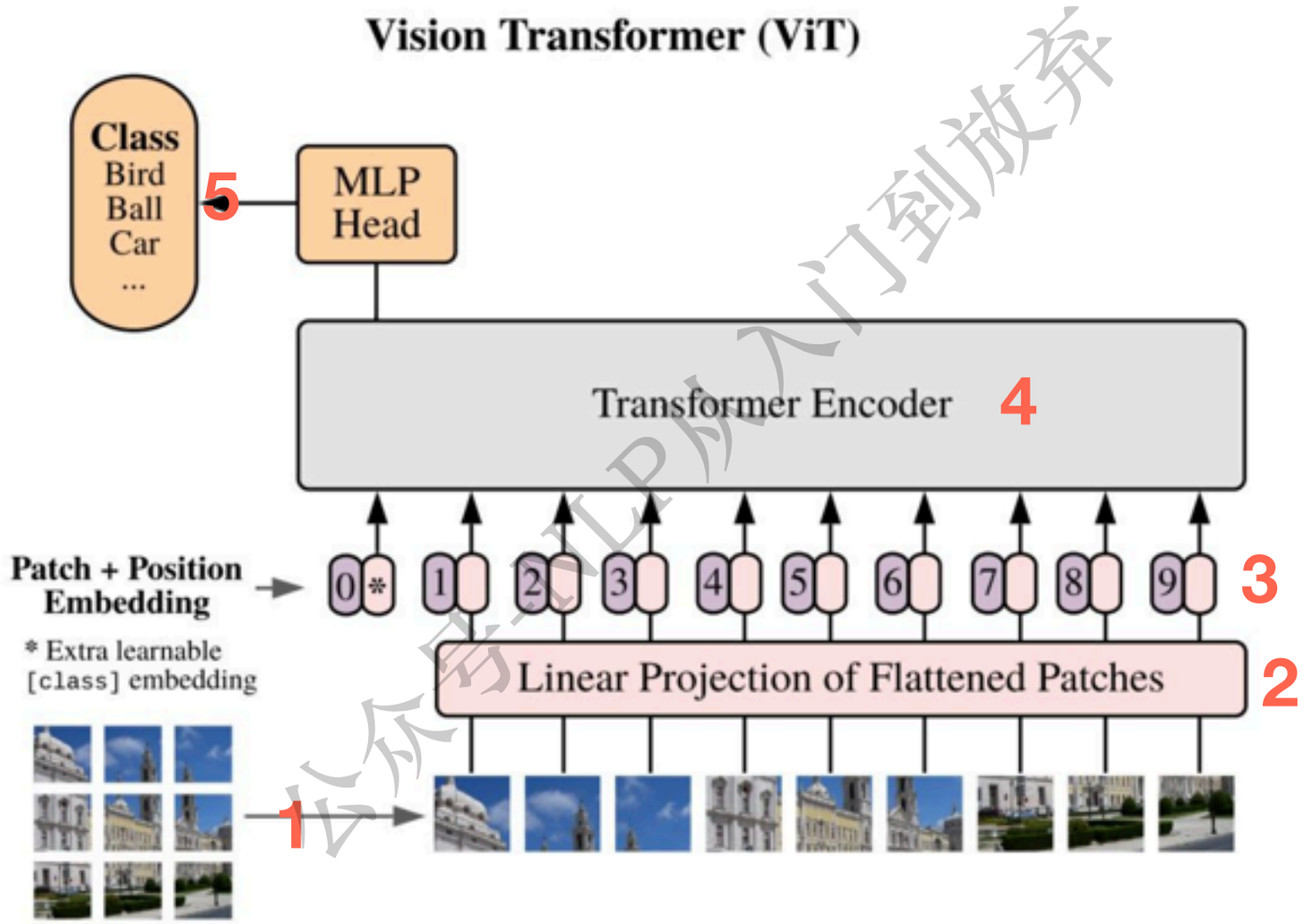


VIT中：



我看了代码，确实有所不同，把Norm提前了/没有pad符号

整体串讲一遍　　**VIT模型架构图：**



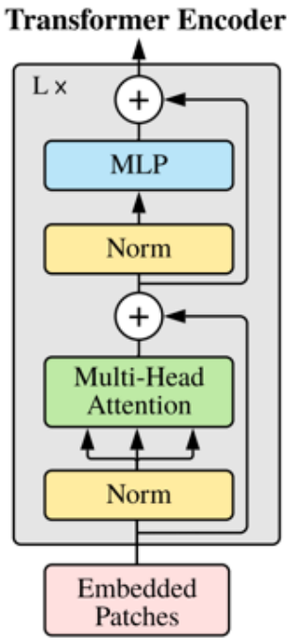
多分类任务

5

768 768 768 768 ... 768

Encoder *N

4



Input embedding 768 768 768 768 ... 768

3

生成位置编码 CLS-768 768 768 768 ... 768

\+ \+ \+ \+ ... \+

拉平+procjction CLS-768 768 768 768 ... 768

2

图片Patch:16*16*3=768

1